VTT Technical Research Centre of Finland

A System Dynamics Model of Data-Driven Precision Medicine Ecosystem

Ylén, Peter; Lähteenmäki, Jaakko; Sorasalmi, Tomi

Published: 01/07/2020

A System Dynamics Model of Data-Driven Precision Medicine Ecosystem

Tomi Sorasalmi, Jaakko Lähteenmäki, Peter Ylén

VTT Technical Research Centre of Finland

Abstract

Biobanking and secondary use of healthcare data enable considerable improvement to healthcare through personalized medicine including innovative therapies, pharmaceuticals, medical technology, and healthcare processes tailored for the needs of individuals. The data-driven precision medicine ecosystem requires contributions from a large variety of stakeholders as well as public investments to boost ecosystem growth. This paper introduces a system dynamics model describing mechanisms for biobank data and biosample accumulation and the related impact to pharmaceutical R&D projects. The preliminary simulation results reflect the important role of public funding of the infrastructure needed for biobank and register data exploitation as well as setting up services to attract more donors. Large public-private precision medicine projects, such as the FinnGen project are in a key role. Further work is needed to improve the ecosystem model precision especially in taking into account the differences between pharmaceutical R&D projects and their data exploitation needs.

Keywords: biobank, system dynamics, healthcare data, precision medicine, business ecosystem, innovation ecosystem

1 Introduction

Biobanking and secondary use of healthcare data enable considerable improvement to healthcare through personalized medicine including innovative therapies, pharmaceuticals, medical technology, and healthcare processes tailored for the needs of individuals. Secondary use of healthcare data enables data to be used for monitoring the quality of provided services, in the development of new services or products and for scientific research. Secondary use is permitted under certain conditions by the existing privacy legislation, in particular, the General Data Protection Regulation (GDPR). Additionally, specific legislation addressing secondary use of health data and biobanking has been implemented in in Finland[1,2].

Based on the legislation on secondary use (2019) the new data permit authority (Findata) was established, with services gradually starting during year 2020. Based on the Biobank Act (2013) ten biobanks have been established and are operating in Finland at the moment: six biobanks covering the university hospital districts and four national biobanks[3]. An important step towards centralised biobank services was the opening of the Fingenious service of the Finnish Biobank (FinBB). Fingenious enables feasibility and access requests to be done for all hospital biobanks and the THL biobank via one joint service.

The new centralized services together with the supporting legislation are targeted to facilitate more efficient and secure use of data of the various local registers, biobanks, centralized Kanta services as

[1] Biobank Act, https://www.finlex.fi/fi/laki/kaannokset/2012/en20120688.pdf
[2] Act on the secondary use of health and social data, https://stm.fi/en/secondary-use-of-health-and-social-data
[3] http://www.bbmri.fi/bbmri-network/finnish-biobanks/

well as the national statistical registers. The services are intended to benefit both the academic community and the industry. Full exploitation of data will require efficient co-operation between all stakeholders including the centralized services (Findata and FinBB), health and social care register controllers, the industry and the providers of research services. The industrial stakeholders include large pharmaceutical and diagnostics companies, food industry, medical device manufacturers and a wide spectrum of SME companies offering research, diagnostics, data management, bioinformatics and software services.

The data-driven precision medicine ecosystem is expected to lead to remarkable economic benefits through new products and business opportunities. However, considerable public investments are needed, e.g. to establish the required infrastructure (data/samples, biobank processes, data application evaluation processes, etc.), to accelerate the R&D of ecosystem companies and to support related academic research. For decision-makers an important question is how to support the data-driven precision medicine ecosystem growth to achieve maximum benefit for the society.

Figure 1 shows the conceptual high-level data-driven precision medicine ecosystem model as identified in the PreMed project (Lähteenmäki *et al.*, 2020). The model is especially targeted to describe the effects of public investments in the ecosystem. The primary outcome measure is the volume of projects - referred as real world data (RWD) projects (Bartlett *et al.*, 2019) - which are based on exploitation of data obtained from biobanks and national registers. The model assumes that willingness of the industry to initiate and execute new RWD projects depends mainly on the availability and value of data, results from past projects (both academic and industry driven) and related services needed in the analysis of data and biosamples (blood or tissue samples).

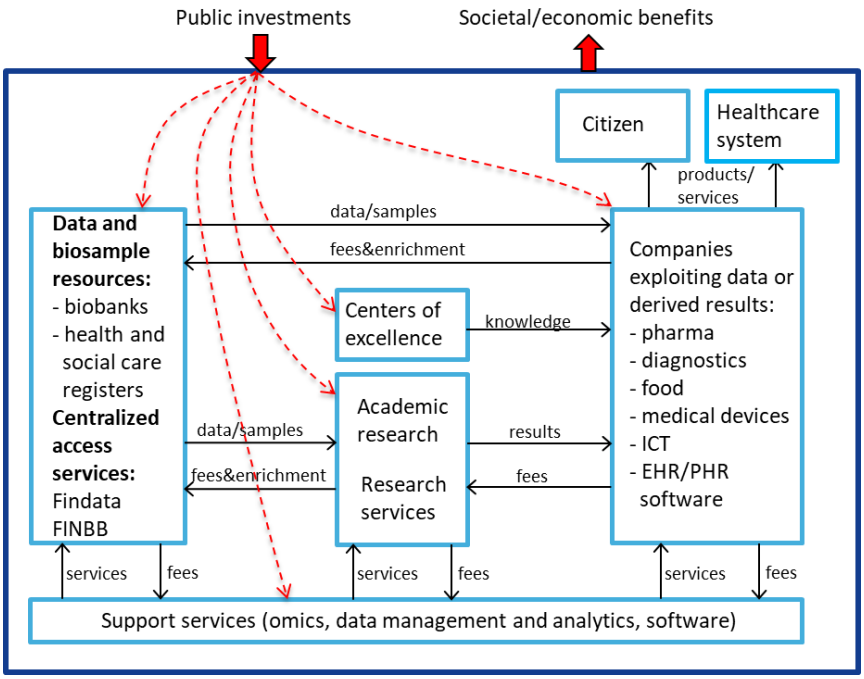

Figure 1. Conceptual high-level ecosystem model showing main stakeholder groups and dependences affecting the exploitation of health data.

The conceptual ecosystem model includes the relevant ecosystem stakeholders and their main connections. The model takes into account the main mechanisms for companies to benefit from secondary use of data: (1) direct use of data by carrying out in-house studies, (2) using external research services, and (3) adopting results from academic studies. In all cases, the use of data

(and/or biosamples) involves fees to be paid and if sample analytics have been carried out in a biobank study, the results need to be returned to the biobank in order to enrich the value of the sample. The ecosystem includes various types of companies either in the role of using data in their R&D activities or in providing support services to facilitate the use of data and samples. Centers of excellence in Figure 1 refer to the Finnish national centers under development in the areas of cancer, neurological disorders, pharmaceutical development and genomics. These organisations are being set up with the objective of promoting research and public-private co-operation in the respective areas.

The conceptual ecosystem model depicted in Figure 1 has been implemented as a simulation model, which enables the analysis of various ecosystem evolution paths affected by different public investment strategies. The model enables the effect of investments to biobanks, Findata, academic research, and support services to be estimated (centers of excellence have been omitted from this version of the model). The simulation model is expected to be valuable for public entities and authorities in providing support for the selection of financing strategies to boost ecosystem growth. Also other stakeholders can find the simulation model useful in increasing the understanding of the ecosystem dynamics.

The approach presented in this paper to model and simulate the evolution of the data-driven precision medicine ecosystem from the perspective of biobanks is novel. Biobanking has been studied mainly from legal and social perspectives, especially concerning the privacy issues regarding healthcare data (Burgess, O'Doherty and Secko, 2008) and (Townend, 2016). Also, the Finnish biobank ecosystem has been studied from a data privacy viewpoint (Soini, 2016). Ecosystem simulation approach, especially system dynamics, has been used to model different kinds of innovation and business ecosystems, for example, in the context of technological innovation ecosystems (Walrave and Raven, 2016; Raven and Walrave, 2018) and in the context of digital service platforms (Ruutu, Casey and Kotovirta, 2017). To our knowledge, ecosystem simulation modelling perspective has not been applied to biobanking.

In this paper we present a model of a data-driven precision medicine ecosystem and initial simulation results of the different future development scenarios of the ecosystem. The aim of the paper is 1) to visualize the interconnections of the different stakeholders of the ecosystem, 2) to study the effects of different sectors of the system on the whole system, 3) to study the long term effects of public investment policies, and 4) study which factors support and which restrict the development of data-driven precision medicine ecosystem. The effects are measured by the number of industrial R&D projects based on exploitation of health data and biosample resources.

2 Modelling the data-driven precision medicine ecosystem

2.1 Model overview

The ecosystem simulation model was constructed based on available open material and information received from representatives of biobanks, healthcare providers, authorities and industry. As already shown in Figure 1, exploitation of health data involves a wide range of companies and public organizations forming a complex system. A simplified simulation model of the ecosystem is presented in Figure 2. The ecosystem model consists of five sectors that interact with each other. The sectors are (a more detailed description of each of the sectors is presented in section 2.2):

1. **Biobanks:**

    - **Donors and donor services:** Donors describes the number of persons who have given a biobank consent for the use of their health care data and samples. Donor services describe the services set up for collecting consents from potential biobank donors. Donor services include also online services for giving and managing the biobank consents and can also be seen as a way of providing information for the donors where their samples and data have been used. It is expected that availability of different kinds of donor services especially in the context of healthcare services will increase interest towards giving biobank consents.

    - **Biobank samples and data:** Describes the amount of samples and data collected from the biobank donors under biobank consent as well as the processes for collecting samples and generating data.

2. **Pharmaceutical R&D:** Describes the research and development activities specifically exploiting health care data and conducted by the pharmaceutical companies. Real world data (RWD) projects are typical examples of such pharmaceutical R&D. The main impacts of the pharmaceutical R&D projects on other parts of the ecosystem are as follows: 1) R&D projects provide service fees to biobanks and Findata, 2) R&D projects return the sample-level analytics results back to biobanks, thus causing sample enriching, and 3) R&D projects use external services from companies and thus enable the operations of support services.

3. **Findata[1]:** Findata provides services for accessing healthcare data for secondary use. Findata focuses on providing availability and data permit services for national register data. Deployment of Findata services is expected to shorten times for accessing register data and thereby increase industrial interest towards data exploitation. Currently, Findata is not covering biobank resources, which need to be applied separately from the biobanks.

    - **Register data:** Describes the amount of register data available. Register data includes patient record data (use of healthcare services, diagnoses, procedures, medication, laboratory, etc.) in local registries and in the centralized Kanta archive as well as data collected in national registers for quality monitoring, statistics and research.

4. **Support services:** Describes the services which the pharmaceutical companies need in their R&D projects and which they are not able (or willing) to do themselves. These services may include different kinds of data analytics, laboratory analyses, software development, etc. Support services consist mainly of private sector companies, but also include the services provided by research centers and other publicly funded organizations. The availability of support services is expected to affect positively in the pharmaceutical companies' interest and ability to conduct R&D projects.

5. **Academic research:** Describes academic research specifically based on the resources of biobanks and health care registers.

[1] https://www.findata.fi/en/data/

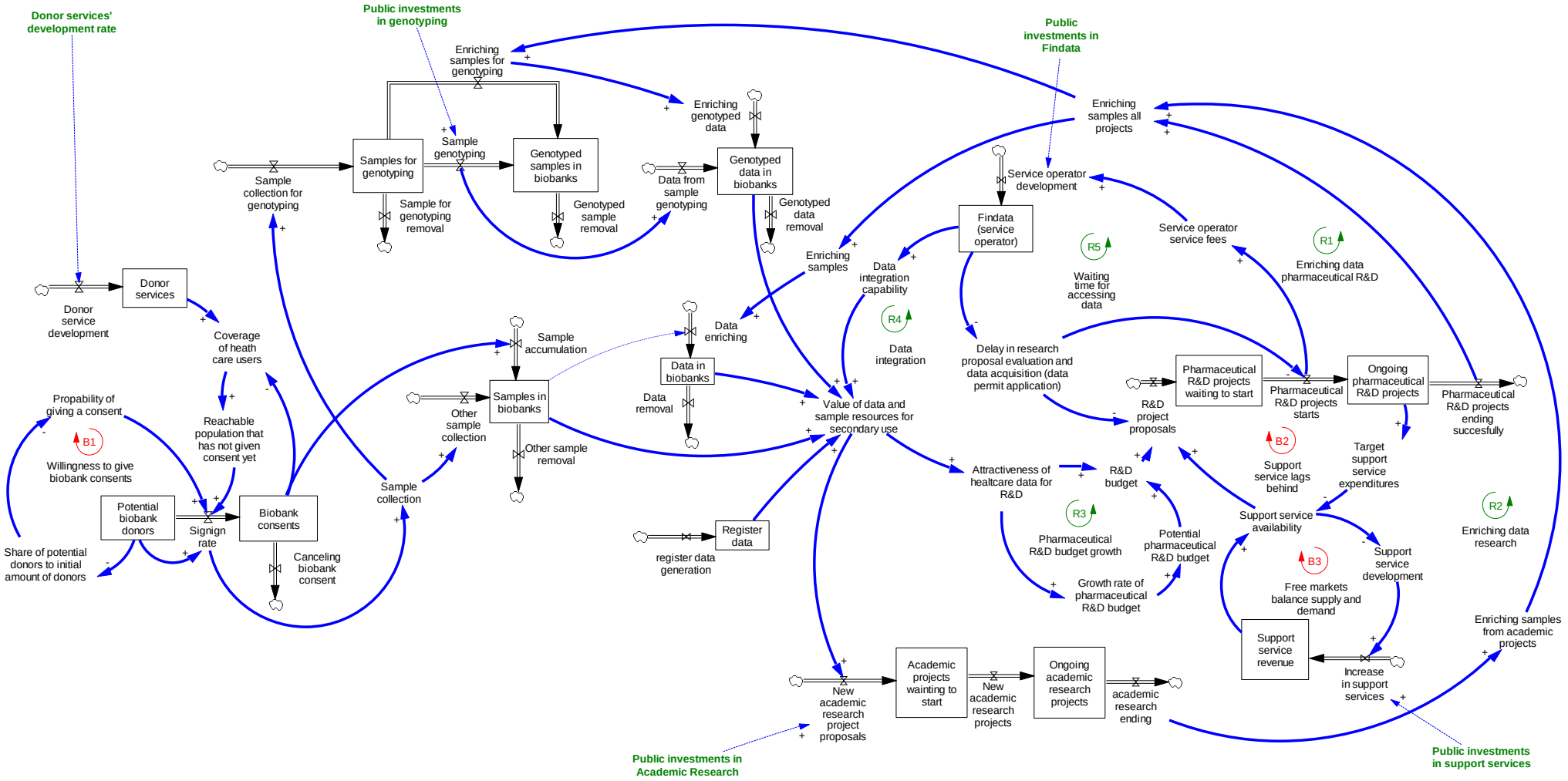

Figure 2: Stylized diagram of the simulation model presenting all model stocks and the main flows and feedback loops.

2.2 Model description

2.2.1 Biobank consents

Figure 3 represents the part of the model that describes the donor services and how biobank consents are collected. Loop B1 causes the number of biobank consents to saturate as the number of biobank consents increase and as the number of potential biobank donors decreases when consents are given and thus the likelihood of finding new biobank donors decreases. The availability of donor services counter-balances the effect of this loop to some extend as long as donor services are developing, but eventually B1 will cause decrease in signing rate.

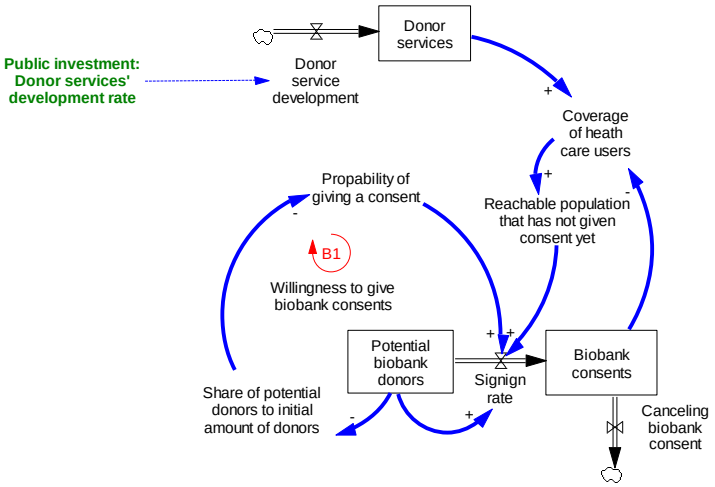

Figure 3: Biobank consent dynamics

Biobank consents has its own dynamics but is not connected to the other parts of the model by feedback loops, that is, nothing in the other parts of the model is feeding back to biobank consents. The number of consents is mainly driven by the personal choice of individuals which can be positively influenced by useful donor services.

2.2.2 Biobanks

The biobank section of the model describes biobank operations from the point of view of generating sample and data resources. It does not cover other operational and financial mechanisms of the biobanks.

Figure 4 presents the model part describing the biobank sample and data generation. Samples are generated along with signed consents. This generates a sample for both *Samples for genotyping* and *Samples in biobanks* stocks. *Genotyped data in biobanks* is generated by direct genotyping funded externally or by enriching samples. *Data in biobanks* is only generated by enriching.

Value of data and sample resources for secondary use is a combination of the following data resources: *Data and samples in biobanks* and *Register data*. Also, the *Data integration capability* affects the amount of data available (described in 2.2.4).

The variable *Samples in biobanks* is assumed to be an inexhaustible source of data, as one sample (blood or tissue) can be used several times for different types of analytics. The variable *Samples for genotyping* on the other hand is based on an assumption that genotyping is done only once per person.

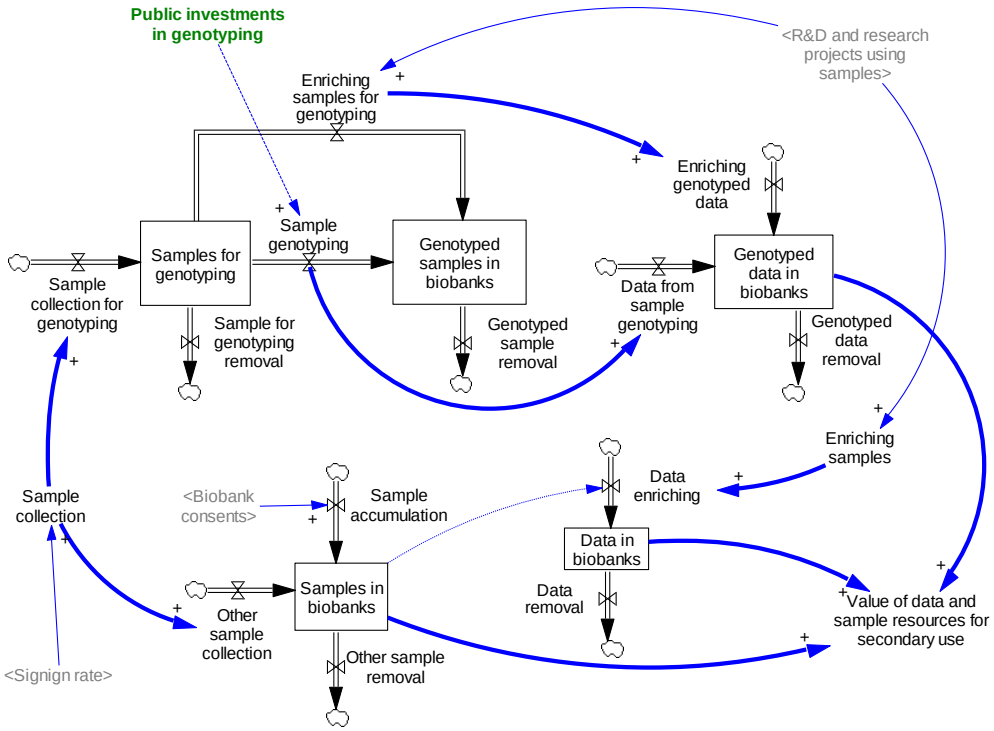

Figure 4: Biobank sample and data generation

Data enrichment is an important feedback from the pharmaceutical R&D and academic research back to biobanks. This feedback is described by loops R1 and R2 in Figure 2. Data enriching, as described by the model, is the only process that is producing data from *Samples in biobanks.* When samples are used in R&D projects, according to the Material Transfer Agreements (MTA) the enriched data has to be returned back to the biobanks. This in turn increases the amount of available data resources for further use and thereby attracts new projects.

It should be noted that *Samples in biobanks* is not directly affecting the enriching data for *Data in biobanks*. This effect comes through projects using the data and then enriching it. However, we have indicated the link by dashed line that the data is produced from the samples.

2.2.3 Pharmaceutical R&D Projects and budget growth

A key objective of the model is to determine what affects the budget available for pharmaceutical R&D project and thus the project starts. Figure 5 presents the variables affecting R&D budget and R&D project proposals.

The main effect comes from the *Value of data and sample resources for secondary use* defined in Section 2.2.2. We have assumed that pharmaceutical R&D is positively affected by the value of health data available for use (affected by the richness of data and samples as well as the possibility to merge data from different sources). We have not taken into consideration the possible effects of the number of donors relative to available data, or the variability of the resources (e.g. blood vs tissue samples), but aggregated the effect of these in the single variable *Value of data and sample resources for secondary use*. We have given different weights to samples and data extracted from the samples, as the data is in more accessible format. Also, genotyped data is given different weight from other data.

The budget reserved by pharmaceutical companies for R&D has a major impact on the project proposals. When the number of R&D projects increases the budget becomes a limiting factor for the projects as

described by loop R3. The variable *Growth rate of R&D budget* is determined by the variable *Attractiveness of healthcare data for R&D*. This loop accounts for the long term growth of the pharmaceutical companies' R&D project volume.

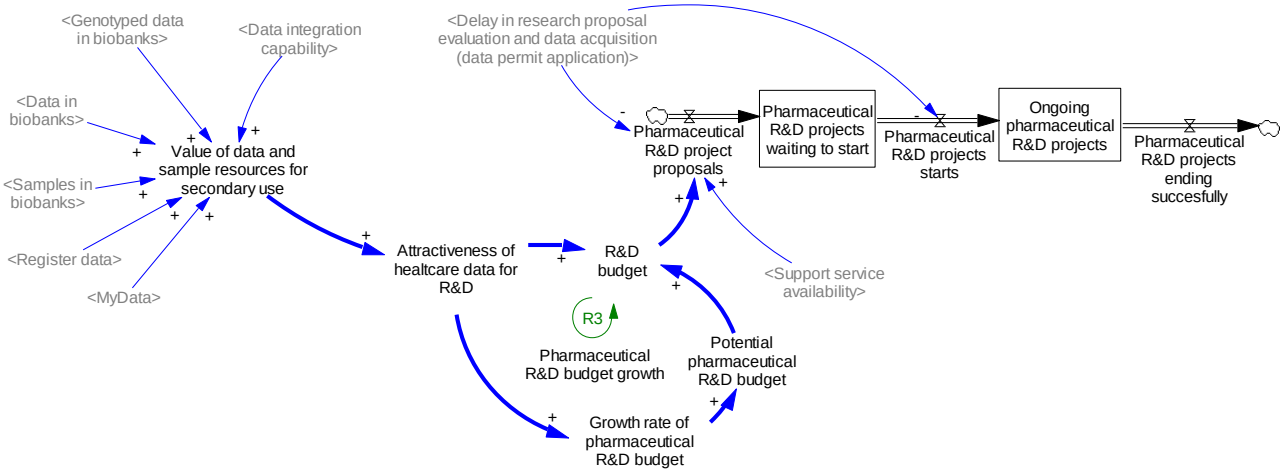

Figure 5: Pharmaceutical R&D

Support service availability has mainly a negative impact: incapability of the support services to match the requirements of the ongoing R&D projects reduces companies´ willingness to initiate new R&D project proposals. The pharmaceutical companies may not be able to carry out their intended R&D projects without appropriate support services. The effect of support services is discussed in more detail in section 2.2.5. The effects of Findata come through the improved application process reducing the needed effort and delay in accessing data as described in section 2.2.4. We have assumed that the application process delay has an effect on how interested the pharmaceutical companies are to initiate new projects.

2.2.4 Findata

Findata is an important stakeholder in promoting the development of the ecosystem. When fully operational, Findata will affect the number of R&D projects mainly in two different ways, shown in Figure 6. First, the evaluation of healthcare data applications will be done in one place (“one-stop-shop”), which is supposed to decrease the evaluation time and labour intensity of the data application process significantly. Secondly, Findata is expected to allow better opportunities for integrating data from various registers.

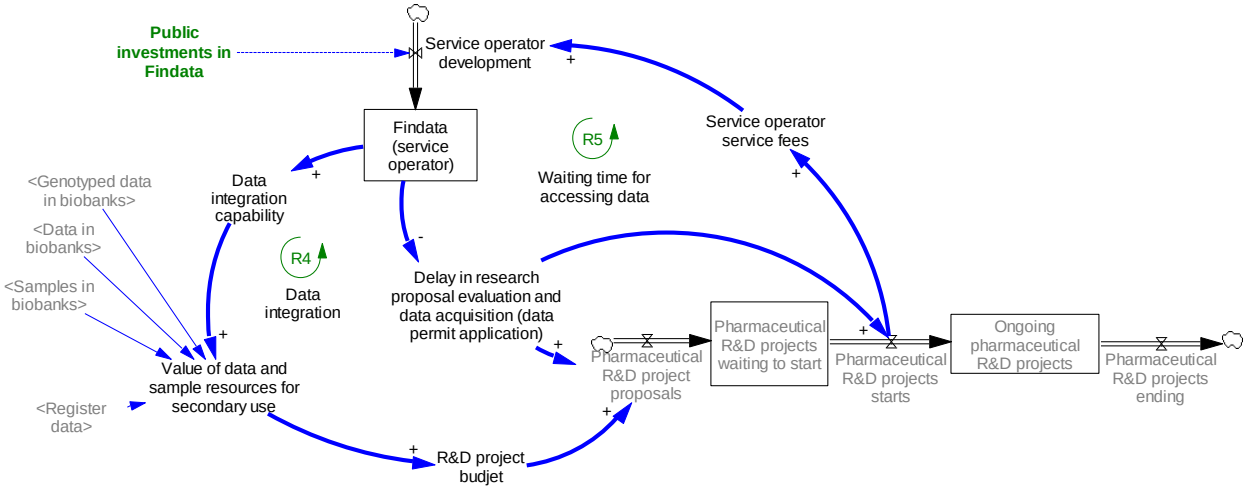

Figure 6: Findata

The financial effect of service fees collected by Findata is not expected to be large. Findata is a public entity supported by the state budget and not intended to create profit. Therefore the importance of the loops R4 and R5 is relatively weak. However, there is a high demand by pharmaceutical companies and research organisations for Findata services, thus motivating and driving its development. This makes public investment in Findata an important decision point in the ecosystem.

Loop R4 describes the possibilities and efficiency of integration of different data sources. The ability to integrate data from various sources, namely from national registers, healthcare service providers' registers and from biobanks, enhances the utility of the data. Data integration capability therefore effect on the value of available data resources for secondary use. The value of data in turn affects the interest of pharmaceutical companies to utilize the data.

Loop R5 describes the delay of Findata's processes affecting the waiting time for data access. Public investments to Findata are used to streamline the data application process and to reduce the application evaluation time.

2.2.5 Support services

The structure of support services is presented in Figure 7. Loop B2 describes how the demand and supply of support services affects the project proposals. As the number of R&D projects is growing relatively fast, it is likely that the supply of services lags behind the demand, depending on the growth strategies (i.e. loop B3) of the service providers. Lack of services is likely to have a negative impact on new project proposals of pharmaceutical companies as already discussed in 2.2.3.

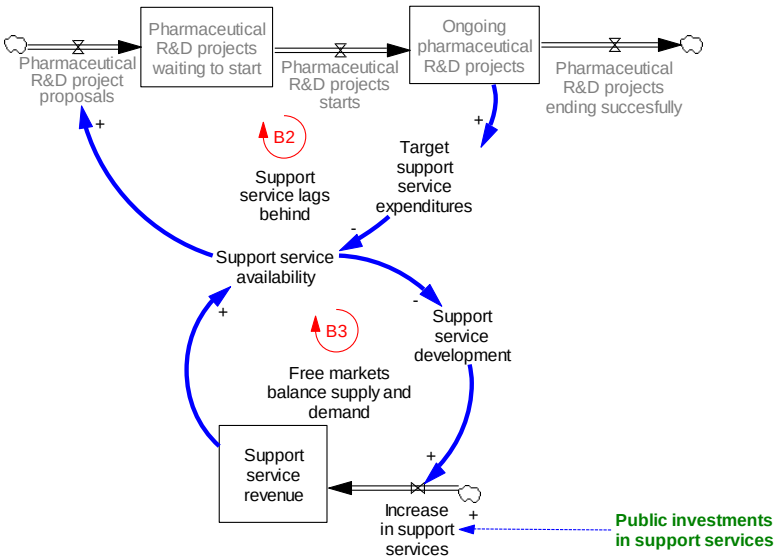

Figure 7: Support services

Loop B3 describes how the supply and demand is balanced, i.e. how the support services react to the growing number of R&D projects. As the number of R&D projects increase, support services try to match this. The faster this loop operates, that is, the smaller the delay for service development is, the faster the service offering of the support service matches the demand of pharmaceutical R&D. Public investment can be used to enhance the growth of the support services and thus balance the supply and demand. This loop, if working slowly, can cause the supply to lag behind the demand, thus hindering the growth of the pharmaceutical R&D through the loop *B2*.

2.2.6 Academic research

In this model, the main impact of academic research is through sample enriching (the same process as in pharmaceutical R&D projects), as research projects have to return the analysed data from the samples back to biobanks.

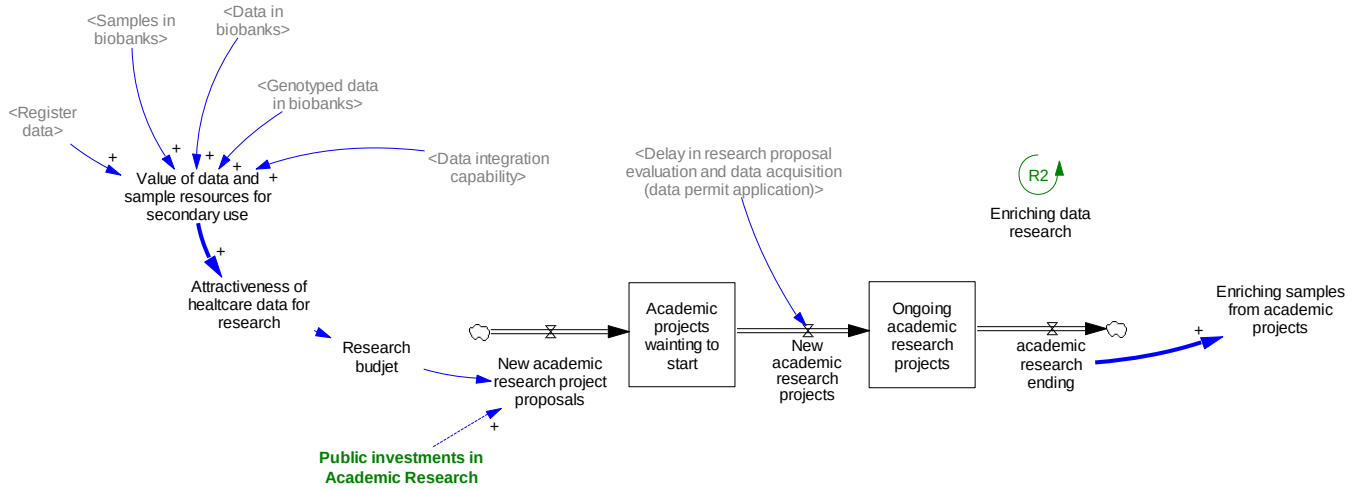

Figure 8: Academic research

3 Simulations

In this chapter preliminary simulation results are presented. The simulation are targeted to gain understanding of the effects of public investments on various sectors of the ecosystem.

The main interest of the simulations was to analyse the amount and evolution of biobank donors, samples and data, and pharmaceutical R&D as well as to study how the public investments (green parameters in the model) affect the evolvement of the ecosystem. In the model, public investments are focused to the following parts of the ecosystem in no particular order: 1) Sample genotyping, 2) Donor services, 3) Findata, 4) Academic research, and 5) Support services:

1. Investments to biobanks are considered from the perspective of genotyping of blood samples, which is mainly financed by research projects such as the ongoing Finngen[5] project.
2. Donor services are considered as a specific area, where targeted investments bring benefits for biobank donors with the objective of encouraging new individuals to contribute as donors.
3. Investments to Findata are provided through the state budget and targeted to setting up the basic processes and information systems to handle data permit applications, connect data resources and to support secure processing of data.
4. Investments to academic research refer to public funding of academic research projects (e.g. via the Academy of Finland) based on retrospective data resources.
5. Investments to support services refer mainly to R&D funding (e.g. via Business Finland) to companies, which provide various services needed in exploiting data and sample resources.

[5] FinnGen, https://www.finngen.fi/en/samples_come_from_biobanks

We have assumed that R&D of pharmaceutical companies do not receive direct public national funding but benefit indirectly from public investments directed to other stakeholders of the ecosystem. Also, the allocation of R&D funding for pharmaceutical companies is seen as an endogenous variable in the model. Thus any public investments (e.g. EU-project financing) for pharmaceutical R&D are considered to be included in the companies' total R&D budgets, and thus not explicitly modelled.

The R&D activities are considered mainly from the point of view of pharmaceutical companies, as it is currently the main sector using biobank data. However, the model could be expanded to include also other emerging businesses utilizing health data.

Table 1 contains parameter values for the scenarios presented. The first and second columns list the names and units of the decision parameters used in different scenarios. The second column (Base) in the table lists parameters reflecting a reference/baseline scenario. The rest of the columns (S0-S5) list parameters representing various alternative scenarios (mark ´-´ means the used value is the same as in Base scenario). If not otherwise indicated, the change is taking place at simulation time (t=0). The simulation start time is year -8, which represents year 2012. Thus, year 0 in the simulation represents year 2020 and year 20 year 2040 respectively.

Table 1: Simulation scenarios

Scenario parameters	Unit	Base	S0	S1	S2	S3	S4	S5
Public investments in genotyping	M€	5	-	-	-	0	5	7
Public investment duration genotyping	Year	3	-	-	-	0	10	20
Public investments in Findata	M€	3	-	-	-	0	6	8
Public investment duration Findata	Year	5	-	-	-	0	6	6
Donor service development rate	Dmnl	0,03	0	0,06	-	0	-	0,06
External biobank consent cancellation ratio	M€	0	-	-	0,15	-	-	-
External biobank consent cancellation duration	Year	0	-	-	3	-	-	-
External biobank consent cancellation start time	Year	0	-	-	5	-	-	-

The initial simulation results are presented in sections 3.1 and 3.2. Total of seven simulations are presented. The first set of simulations concentrate on acquiring biobank consents by developing donor services. The second set of simulations present cases showing an optimistic and pessimistic scenarios. Scenarios are compared to baseline (Base) scenario.

3.1 Base case, S00, S01, and S02

Scenarios S00, S01, and S02 are presented in Figure 9, Figure 10, Figure 11, and Figure 12, showing the effect of two different donor service development rate (proportional yearly growth) and one scenario where people are withdrawing their consent (for some external reason, e.g. as a result of concerns to personal information security) and these are compared to baseline scenario. In scenario S00 the development of donor services is stopped at year 0. In scenario S01 donor services are developed double as fast as in baseline scenario. Scenario S02 presents an external biobank consent cancellation occurrence starting at year 5, lasting 3 years, and cancellations being 15% annually. For detailed description see Table 1.

Figure 9 presents the development status of donor services. The development status is described by a proportional value between 0 and 1. Maximum value (1) corresponds to the target situation where donor services are largely available for giving and managing biobank consents as well as for getting personal information about sample usage.

Figure 10 presents the effect of donor service trajectories and consent cancellation on the development of biobank consents. The dynamics of biobank consent follow S-shaped growth curve as can be seen. The limiting factor is the amount of potential donors, which causes saturation as the probability of new consents decline when a major part of the population has already given the consent. The faster donor service development rate shown in scenario S01 causes faster consent accumulation, although the saturation causes the base scenario to catch up later. The lack of donor service development in scenario S00 causes the consents to saturate very fast. The consent cancellation scenario S02 causes a significant drop in consents and after that the growth trajectory is not as steep as in the base scenario (the consent cancellation has also an effect on the amount of potential donors, as it is easier not to give consent at all than to cancel it).

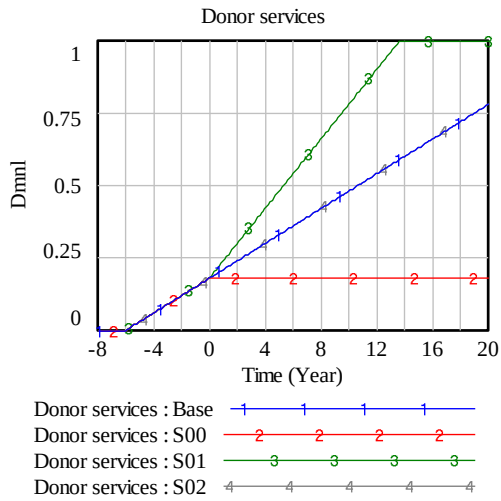

Figure 9. Donor services.

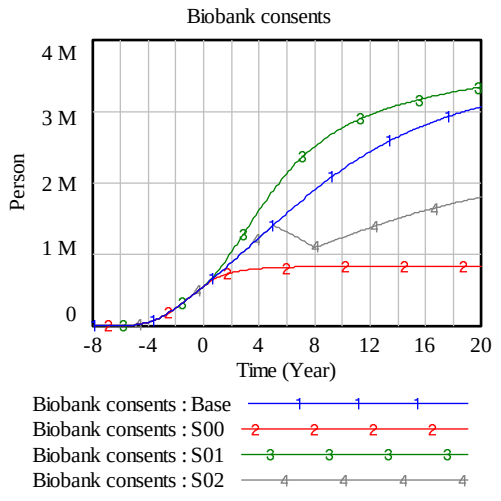

Figure 10. Biobank consents.

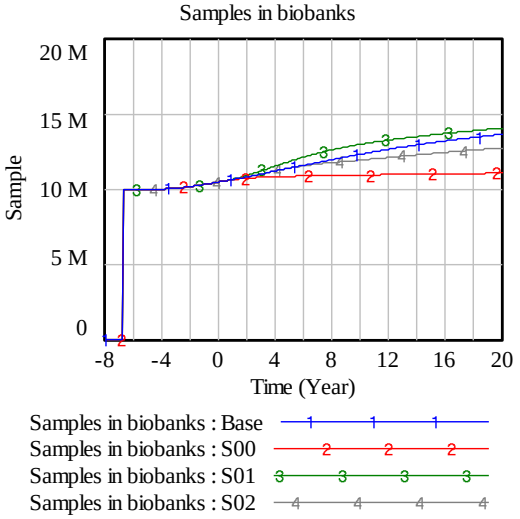

Figure 11. All samples in biobanks.

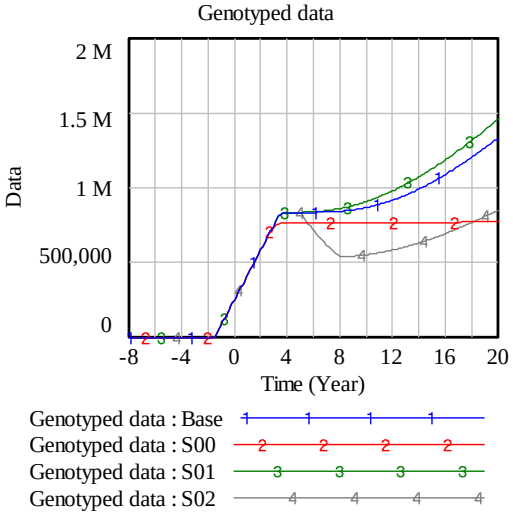

Figure 12. Genotyped data.

Figure 11 presents the evolution of sample resources in biobanks. The amount of samples starts from zero, which describes the situation when the biobanks were not yet founded. When the biobanks were founded, a total of approximately 10 million samples (samples from 4 million persons[6]) were transferred to the

[6] http://www.bbmri.fi/fi/researchers-tools/sample-counter/

biobanks. After the initial sample dump the samples start to flow into the biobanks as consents are given, one sample per one consent. Also, when the consent has been given, samples accumulate slowly as additional biobank samples are taken from the donor in the context of healthcare services. Because of the initial dump of samples, the new samples collected together with new consents have a relatively small effect on the total amount of samples in biobanks.

Figure 12 presents the accumulation of genotyped data. Genotyped data accumulation starts in the model roughly at time -2, after the biobanks have started collecting consents. In base scenario and scenarios S01-S03 public funding for genotyping is stopped at year 3, after which enriching by R&D and academic projects is the only process producing genotyped data. In S00 the genotyping is stopped already before year 3, because of the slow development of consents all samples are already genotyped (not shown in the figures). For base scenario, S00, and S01 enriching has a significant effect on genotyped data in the long run, although it starts slowly. The effect of consent cancellation, S02, has also a significant effect on genotyped data, as it needs to be deleted after consent cancellation.

3.2 Baseline, S03, S04, S05

The following figures present the simulation results of four different scenarios: 1) Base: Baseline scenario, 2) S03: no investments, 3) S04: moderate additional investments, and 4) S05: heavy additional investments. For detailed description see Table 1.

Figure 13 presents the development of biobank consents (for more details see discussion related to Figure 10 in section 3.1). Figure 14 presents the development of samples in biobanks (for more details see discussion related to Figure 11 in section 3.1).

Figure 16 presents the accumulation of genotyped data and Figure 15 the samples available for genotyping. In scenario S03 funding for genotyping is stopped at year 0, so after that only enriching causes slow increase in genotyped data. In scenario S04 public funding for genotyping is continued until year 10. In scenario S05 increased funding for genotyping is continued for 20 years. However, after year 10 the genotyping drops because there is not enough new samples to be genotyped as can be seen in Figure 15. In this situation all new samples are proactively genotyped after taken from the donor.

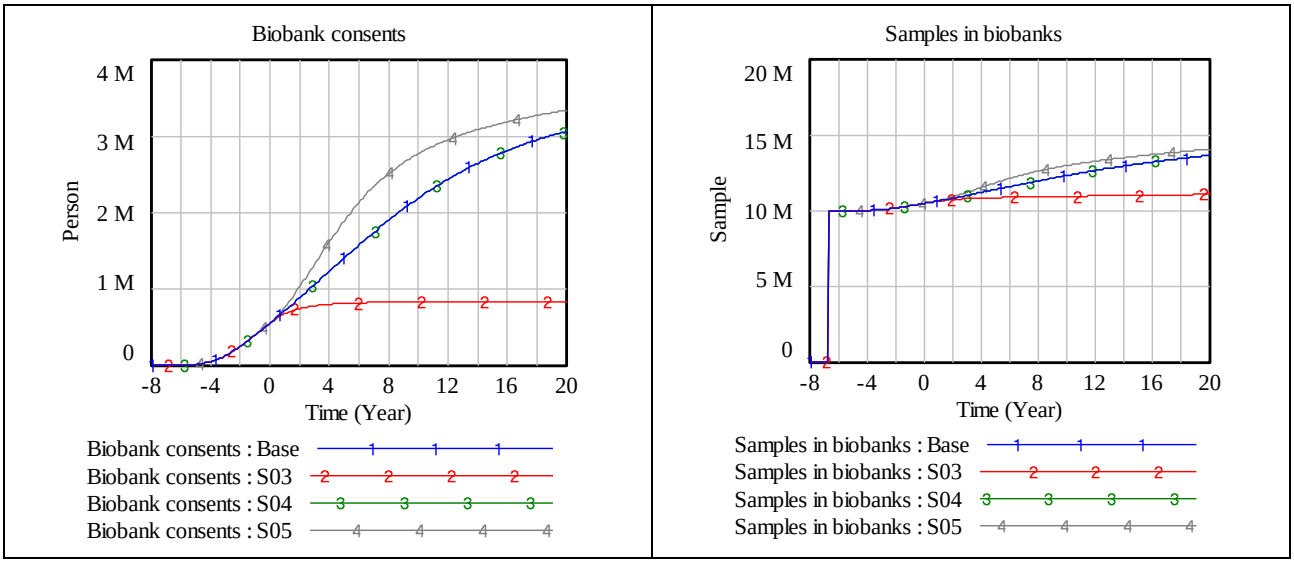

Figure 13. Biobank consents.

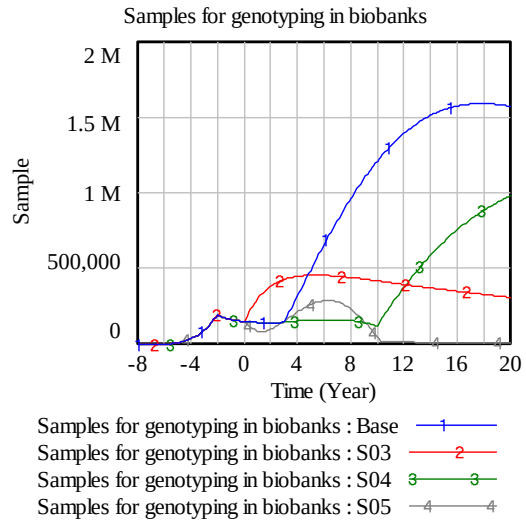

Figure 15. Samples for genotyping.

Figure 14. All samples in biobanks.

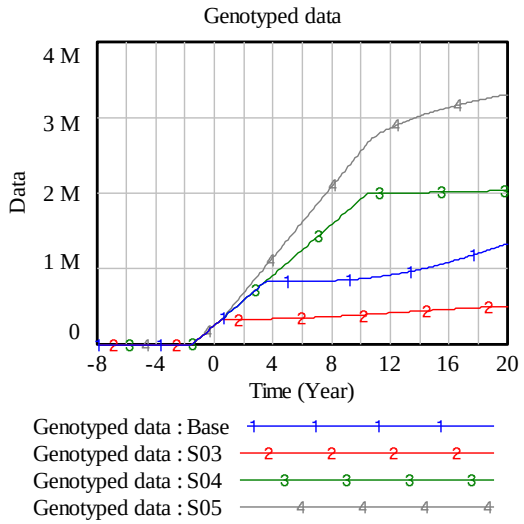

Figure 16. Genotyped data.

Figure 17 and Figure 18 present the development of Findata and the related integration capability respectively. Both Findata progress and data integration capability are described by a proportional value between 0 and 1. For Findata, the maximum value (1) corresponds to the target level of services and capabilities where data is largely available via Findata services. For integration capability maximum value (1) means that data can be efficiently integrated from all health data registers for secondary use. Integration capability starts from 0.5, because already prior to Findata establishment integration has been possible to some extend.

Figure 19 presents the value of data and sample resources perceived by the pharmaceutical companies and academic research community. Scenario S03 presents the effect of inadequate amount of genotyped data and poor integration capability combined. Base, S04, and S05 scenarios present the combination of increasing amount of genotyped data, samples, and integration capability, from which S05 results in highest value of data and samples.

Figure 20 presents how the value of data and sample resources impact R&D project starts. This is also affected by other variables, but the value of data and sample resources has a significant effect. Base, S04, and S05 scenarios are converging towards each other the further the simulation runs, even though this effect is not seen in **Error! Reference source not found.**. This is because the effect of value of data and sample resources has saturated. The model assumes that with enough sample and data resources the additional resources will have a diminishing effect on the value of data, so at some point more data is not going to increase the value of data any more, thus the saturation.

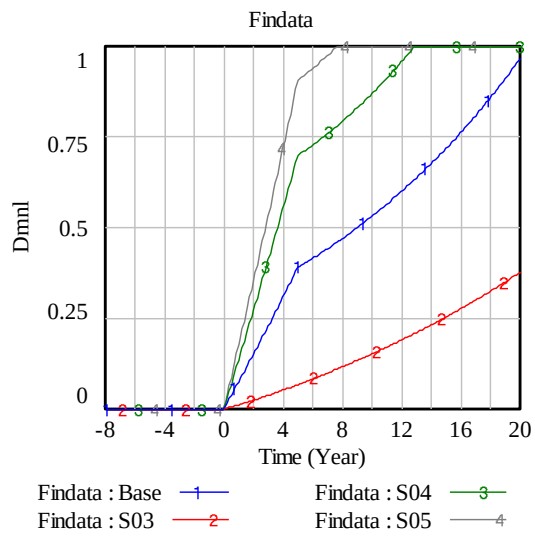

Figure 17. Findata progress

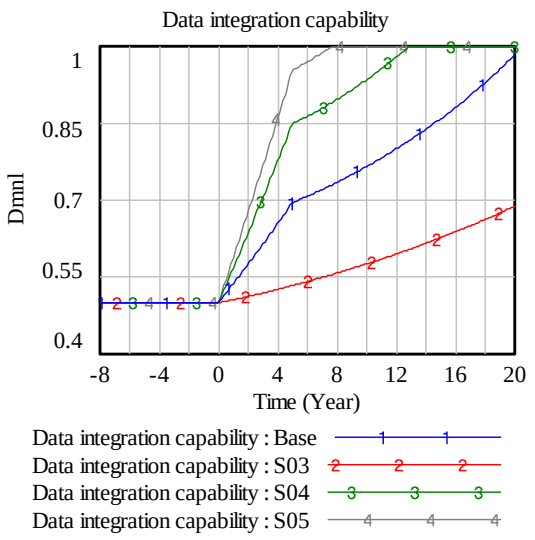

Figure 18. Data integration capability

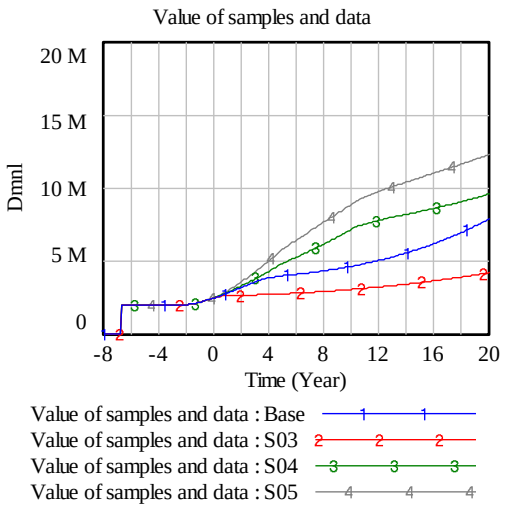

Figure 19. Value of data and samples for secondary use

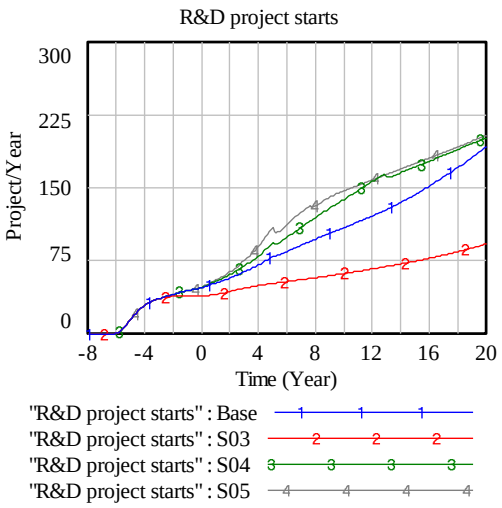

Figure 20. Pharmaceutical R&D project starts

4 Discussion

4.1 Model limitations

As the ecosystem under study is extremely complicated it has been possible to include only selected aspects of it in the model. The ecosystem mechanisms are modelled by simplified approximations which still need further development. Short discussion on the main limitations is given below.

An important limitation of the model is the link between sample and data resources and the pharmaceutical companies' decision to start a new R&D project exploiting data resources. We do not know how companies value different types of data and sample resources and how large data and sample sets are needed. The different types of samples could be separately modelled. In the current model version, new samples have relatively small effect in increasing the existing sample resources of the biobanks (see Figure 11). It may be required to model different sample types to get more realistic results, as the value of samples is expected to vary significantly between sample types. It is also evident, that there are various

types of pharmaceutical R&D projects where data can be exploited and the requirements concerning data and samples are highly variable. Further development of the model could include division between project types including, for example drug discovery, clinical trial support (subject identification and stratification), pharmacovigilance and impact assessment, as this may help to define the link between pharmaceutical R&D and perceived value of data.

Modelling of the financial operation of biobanks has been omitted in the model. Such modelling would have required detailed information of the incomes and expenditures of biobanks, which has not been available for the project. Consequently, the effect of the service fees from R&D and academic projects to biobanks have not been included in the current model. In the future, these service fees are expected to have a growing effect on the biobank economy and to the accumulation of the sample and data resources.

For the pharmaceutical companies and support services the availability of skilled personnel could become a bottleneck for future growth. Especially the lack of data analysts is seen potential hindrance to the ecosystem. This effect has been omitted from this version of the model.

Academic research provides results with potential impact for new product development and related industry driven research projects. National centers of excellence (Genome Center, National Cancer Center, Drug Development Centre, Neurocenter Finland) are expected to have a positive effect to initiation of industrial R&D projects based on academic research results. However, the current version of the model does not explicitly include the effect of academic research and the centers of excellence on industrial R&D projects. The link between academic results (e.g. publications) and industrial R&D is not clear and its modelling requires further effort. The current model version still includes the effect of academic research on sample and data generation and the related indirect impact to R&D projects.

More background information is being collected in order to improve the model structure and parameter values.

4.2 Model validation

Model validation has been an ongoing process throughout the model development. The structure of the model is based on expert knowledge as well as interview with different stakeholders of the ecosystem and the model has been exposed to feedback of different stakeholders. Also, we have used several of the standard model validation tests (Sterman, 2000). For example, we have used partial model testing to see how the model behaves when some parts of the model are removed (e.g. academic research and support services) and sensitivity analyses to study the sensitivity of the model with respect to different parameter values. We have also conducted out extreme condition tests to ensure realistic behaviour under different extreme conditions.

Figure 21 and Figure 22 present biobank consents and the total number of research projects using biobank data. The data consists only a couple years of data, and thus the value of fitting the model to the data is quite limited in validating the model. However, the model can reproduce the data as can be seen.

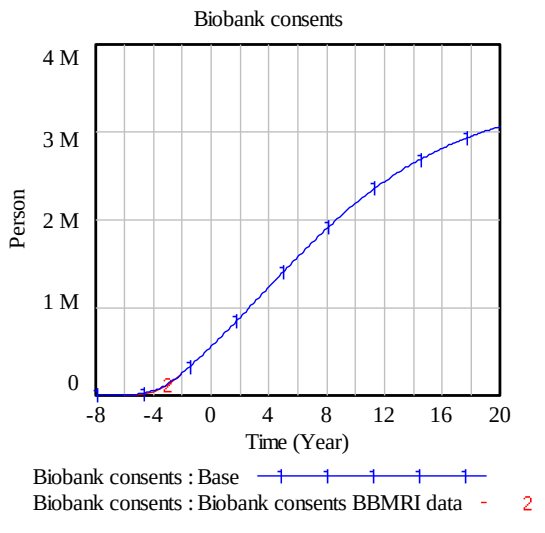

Figure 21. Biobank consents.

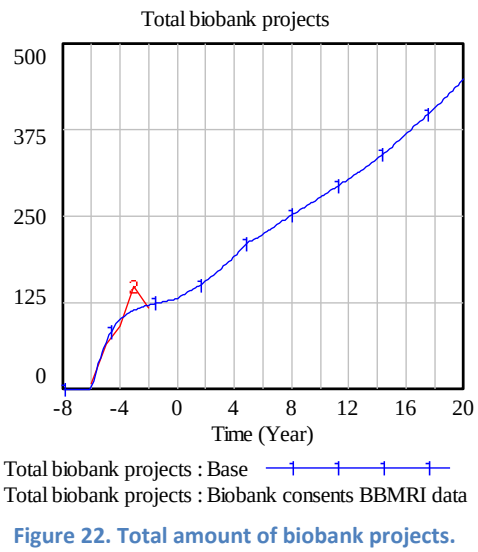

Figure 22. Total amount of biobank projects.

4.3 Discussion on results

The growth of the ecosystem is largely based on public investments that lead to increased amount of samples and data in biobanks. The simulation results indicate a favourable growth trend as long as public investments continue to support the accumulation of biobank consents, samples and data. Along with sample and data collection, Findata and the related infrastructure needed for integrating data is expected to have an important effect on the ecosystem. There are many internal and external aspects that can slow down the rapid development of the ecosystem and the public investment policy will have a remarkable impact on the growth trajectory of the business ecosystem, as can be seen from the preliminary simulation results presented in this paper.

The services for biobank donors are still scarce. However, in line with the GDPR, register controllers would need to provide information for donors concerning the use of biobank samples and patient data. Consequently, there are currently ongoing initiatives towards setting up services for biobank donors, and new services are expected to gradually become available in 2021.

5 Conclusions

The paper describes a system dynamics model for data-driven precision medicine ecosystem, especially focusing on the analysis of biobank data and biosample accumulation and the related impact to pharmaceutical R&D projects. The model describes biobanks, pharmaceutical R&D projects, academic research, registry data access services (Findata) and R&D support services from the point of view of accumulation of data and sample resources. The main output and indicator of ecosystem growth is the number of initiated R&D projects with data exploitation as a key element. The system dynamics model allows the simulation of various factors affecting the ecosystem growth. For example, the disturbance of ecosystem growth due to large-scale cancellation of biobank consents and lack of donor services can be analysed.

Preliminary simulation results are presented concerning the accumulation of biobank consents, samples, data, genotyped data as well as the evolution of Findata services, related capabilities for data integration and the initiation of pharmaceutical R&D projects. While the accumulation of data and sample resources is always limited by the population, a significant improvement can be achieved by public investments. These are needed in provision of donor services, sample genotyping, and in building infrastructures for data access and integration (Findata services). Large public-private precision medicine projects, such as the FinnGen project, are an important channel for focusing public funding to stimulate accumulation of sample and data resources.

As the data-driven ecosystem is extremely complicated it has been possible to include only selected aspects of it in the model. The presented simulation results should be considered only indicative showing overall trends and dependencies. The model development activity continues with the objective of achieving a better coverage of the mechanisms affecting ecosystem growth. Further work is needed for example to take into account the differences between pharmaceutical R&D projects and their data exploitation needs. Also extension to R&D projects beyond the pharmaceutical industry is a potential target for further development.

6 References

Bartlett, V. L. *et al.* (2019) 'Feasibility of Using Real-World Data to Replicate Clinical Trial Evidence', *JAMA network open*. American Medical Association, 2(10), pp. e1912869–e1912869. doi: 10.1001/jamanetworkopen.2019.12869.

Burgess, M., O'Doherty, K. and Secko, D. (2008) 'Biobanking in British Columbia: discussions of the future of personalized medicine through deliberative public engagement', *Personalized Medicine*, 5(3).

Lähteenmäki, J. *et al.* (2020) *Data-driven precision medicine ecosystem - PreMed phase 2 report*. Espoo. Available at: https://cris.vtt.fi/en/publications/data-driven-precision-medicine-premed-phase-2-report.

Raven, R. and Walrave, B. (2018) 'Overcoming transformational failures through policy mixes in the dynamics of technological innovation systems', *Technological Forecasting and Social Change*, (May). doi: 10.1016/j.techfore.2018.05.008.

Ruutu, S., Casey, T. and Kotovirta, V. (2017) 'Development and competition of digital service platforms: A system dynamics approach', *Technological Forecasting and Social Change*, 117, pp. 119–130. doi: 10.1016/j.techfore.2016.12.011.

Soini, S. (2016) 'Biobanks as a Central Part of the Finnish Growth and Genomic Strategies: How to Balance Privacy in an Innovation Ecosystem?', *The Journal of Law, Medicine & Ethics*, 44(1), pp. 24–34.

Sterman, J. D. (2000) *Business Dynamics: Systems Thinking and Modeling for a Complex World*. New York, NY, USA: McGraw-Hill Companies.

Townend, D. (2016) 'EU Laws on Privacy in Genomic Databases and Biobanking', *The Journal of Law, Medicine & Ethics*, 44(1), pp. 128–142.

Walrave, B. and Raven, R. (2016) 'Modelling the dynamics of technological innovation systems', *Research Policy*. Elsevier B.V., 45(9), pp. 1833–1844. doi: 10.1016/j.respol.2016.05.011.